AF
JFW

S&H Form: (2/01)

Docket No.: 1668.1021

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE
BEFORE THE BOARD OF PATENT APPEALS AND INTERFERENCES

In re the Application of:

Ophir FRIEDER et al.	
Serial No. 09/629,175	Group Art Unit: 2171
Confirmation No. 4562	
Filed: July 31, 2000	Examiner: Le, Uyen T

For: SYSTEM FOR SIMILAR DOCUMENT DETECTION

RESPONSE TO NON-COMPLIANT APPEAL BRIEF UNDER 37 C.F.R. § 41.37

Commissioner for Patents
PO Box 1450
Alexandria, VA 22313-1450

Sir:

This is in response to a Notification of Non-Compliant Appeal Brief dated June 6, 2006 and having a period for response set to expire on July 6, 2006. The Notification indicates that the headings of the various sections do not comply with 37 CFR § 41.37. The headings have been revised to be identical to those mentioned in 37 CFR § 41.37.

The claims appendix was presented under Roman numeral X, which has been corrected to Roman numeral IX. It appears that the Examiner wishes for the Roman numeral to be changed to - -VII - -. However, there is no such requirement in 37 CFR § 41.37. This rule provides "The brief shall contain the following items under appropriate headings and in the order indicated in paragraphs (c)(1)(i) through (c)(1)(x) of this section." The rules use (i) through (x) simply to enable the presentation of a list. There is no requirement that these Roman numerals be used. In fact, if there were such a requirement, then the claims appendix would need to be labeled (c)(1)(vii). Clearly, the rules never intended such a pendantic hypertechnical interpretation. Accordingly, the Roman numerals have been substantially maintained as originally filed.

Other than the changes mentioned above, what follows is identical to the Appeal Brief originally filed on April 10, 2006.

******

In a Notice of Appeal filed December 9, 2005, the Applicants appealed the Examiner's September 9, 2005 Office Action finally rejecting claims 1-29, 44-48 and 50-57. Therefore, Submitted herewith are Appellants' Brief, the requisite fee set forth in 37 C.F.R. § 41.20(b), the petition for extension of time and the extension of time fee.

I. REAL PARTY IN INTEREST (37 CFR § 41.37(c)(l)(i)

The real party in interest is Alion Science and Technology, Corporation, the assignee of the subject application.

II. RELATED APPEALS AND INTERFERENCES (37 CFR § 41.37(c)(l)(ii))

Appellants, Appellants' legal representatives and the assignee are not aware of any other appeals or interferences which will directly affect or be directly affected by, or have a bearing on, the Board's decision in the pending appeal.

III. STATUS OF CLAIMS (37 CFR § 41.37(c)(l)(iii))

Pending appealed claims 1-28, 44-48, and 50-57 have been rejected. Claims 29 and 34-43 have been canceled. Finally, claims 30-33, 49, and 58-62 have been allowed.

IV. STATUS OF AMENDMENTS (37 CFR § 41.37(c)(l)(iv))

A first Amendment After Final was filed on November 14, 2005. In an Advisory Action dated November 29, 2005, the Examiner refused to enter the first Amendment After Final. A second Amendment After Final was filed on April 6, 2006. The second Amendment After Final simply cancels one claim. Both 37 C.F.R. § 1.116 and 37 C.F.R. § 41.33 permit entry of amendments to cancel claims. A copy of the second Amendment After Final is enclosed, together with the patent office date-stamped postcard.

The claims as they appear in Section X (Claims Appendix) are the claims as amended by the second Amendment After Final. That is, Section X (Claims Appendix) assumes that the second Amendment After Final has been entered.

V. SUMMARY OF CLAIMED SUBJECT MATTER (37 CFR § 41.37(c)(l)(v))

A. Claim 1

Independent claim 1 recites a method for detection of similar documents, for example the process depicted in Figure 1. *See* specification page 7, line 1 through page 13, line 27.

Claim 1 further recites obtaining a document, disclosed in the embodiment of Figure 1 as block 1. For example, the document may be obtained via a computer network, a computer readable medium, or an optical character recognition document of a scanned paper document. *See* specification page 10, lines 25-30.

Claim 1 further recites filtering the document, disclosed in Figure 1 as block 2. For example, the filtering step is based on parts of speech. *See* specification page 12, lines 23-30. Alternatively or additionally, the filtering step may be based on collection statistics relating to a number of occurrences of words or phrases in the document *See* specification page 11, lines 1-7, for example.

Claim 1 further recites sorting the filtered documents according to a predetermined ranking, disclosed Figure 2 as block 18. For example, retained words may be arranged in Unicode ascending order. *See* specification, page 14, lines 29 through page 15, line 6.

Claim 1 further recites generating a single tuple for the filtered document, disclosed in Figure 1 as block 5. For example, the tuple comprises the document identifier of block 3 and the hash value determined in block 4. *See* specification, page 16, lines 3-6.

Claim 1 further recites comparing the tuple generated with a plurality of tuples representing the contents of a document storage structure, disclosed in Figure 1 as blocks 6-9. For example, the tuple generated in block 4 is compared to a hash table that represents a plurality of document stored in a document storage structure. *See* specification, page 16, lines 7-16.

Claim 1 further recites determining if the document's tuple is clustered with another tuple in the document storage structure, thereby detecting if the document is similar to another document represented by the other tuple in the document storage structure. This feature corresponds with block 7 in Fig. 1 and is described in the specification at page 17, lines 22-30, for example. See also blocks 8 and 9 in Fig. 1 and specification page 18, lines 1-12.

Claims 50 and 51

With regard to independent claims 50 and 51, similar documents are detected, for example by the process depicted in Figure 1. See specification page 7, line 1 through page 13, line 27. To detect similar documents, a document is obtained, as disclosed in the embodiment of Figure 1 as block 1. For example, the document may be obtained via a computer network, a computer readable medium, or an optical character recognition document of a scanned paper document. See specification page 10, lines 25-30.

The document is filtered to eliminate tokens based on parts of speech and to obtain a filtered document. See Figure 1, block 2 and specification page 12, lines 22-30. A single tuple is generated for the filtered document, as disclosed in Figure 1 at block 5. Page 16, lines 3-6 of the specification describe that the tuple may comprise the document identifier of block 3 and the hash value determined in block 4.

The tuple generated is compared with a document storage structure comprising a plurality of tuples, each tuple representing one of a plurality of documents. See Figure 1, blocks 6-9 and specification, page 16, lines 7-16, which describe that the tuple generated in block **4** is compared to a hash table that represents a plurality of document stored in a document storage structure.

It is determined if the tuple for the filtered document is clustered with another tuple in the document storage structure. By doing this, a similar document, represented by another tuple in the document storage structure, can be detected. See blocks 7-9 in Figure 1 and specification page 17, line 22 through page 18, line 12, for example

VI. GROUNDS OF REJECTION TO BE REVIEWED ON APPEAL (37 C.F.R. § 41.37(c)(l)(vi))

The issue presented for appeal is whether claims 1-28,44-48, and 50-57 are distinguishable over U.S. Patent No. 6,240,409 to Aiken.

VII. ARGUMENT (37 C.F.R. § 41.337(c)(l)(vii))

Rejection of Claims 1-29, 44-48, 50-57 Under 35 U.S.C. § 103

A. The Law Regarding the Obviousness Issues related by the Examiner

Under Graham v. John Deere Co.,383 U.S. 1,148 U.S.P.Q. 459 (1966) the scope and content of the prior art are to be determined, the differences between the prior art and the claims at issue are to be ascertained and the level of skill in the art is to be ascertained. Against this background the obviousness of the subject matter is determined.

Obviousness cannot be established by combining the teaching of the prior art to produce the claimed invention, absent some teaching or suggestion supporting the combination. Under section 103, teachings of references can be combined only if there is some suggestion or incentive to do so. (see ACS Hospital Systems, Inc. v. Montefiore Hospital, 221 USPQ 929, 932, 933 (Fed. Cir. 1984))

The prior art must not only suggest the desirability that the teachings of references be combined but must also suggest the desirability of the modifications in the manner proposed by the Examiner as well as the results to be achieved (see Ex parte Costa,211 U.S.P.Q. 636(P.O.Bd.App.1978), ACS Hospital Systems, Inc. v. Montefiore Hospital,732 F.2d 1572,221 U.S.P.Q. 929(Fed.Cir.1984), In re Gordon,733 F.2d 900,221 U.S.P.Q. 1125(Fed.Cir.1984), Lear Siegler v. Aeroquip Corp.,733 F.2d 881,221 U.S.P.Q. 1025(Fed.Cir.1984) and Diversitech v. Century Steps,850 F.2d 675,7 U.S.P.Q.2d 1315(Fed.Cir.1988)).

To support a finding of obviousness based on a single reference, the single reference must suggest the desirability of modifying its disclosure as needed to accomplish the invention (see In re Gordon,733 F.2d 900,221 U.S.P.Q. 1125(Fed.Cir.1984), Schneck v. Gordon,713 F.2d 782,218 U.S.P.Q. 699(Fed.Cir.1984) and Cooper v. Ford,748 F.2d 677,223 U.S.P.Q. 1286(Fed.Cir.1984)).

The Examiner bears the initial burden of establishing a prima facie case of obviousness. See In re Rilckaert, 9 F.3d 1531, 1532, 28 USPQ2d 1955, 1956 (Fed. Cir. 1993). A prima facie case of obviousness is made by presenting evidence that the "reference teachings would appear to be sufficient for one of ordinary skill in the relevant art having the references before him to make the proposed substitution, combination or other modification." In re Lintner, 458 F.2d 1013, 1016, 173 USPQ 560, 562 (CCPA 1972); In re Lalu, 747 F.2d 703, 705, 223 USPQ 1257, 1258

(Fed. Cir. 1984). It is incumbent on the Examiner to state how and why the teachings of the references would have been combined. "If examination at the initial stage does not produce a prima facie case of unpatentability, then without more the applicant is entitled to grant of the patent." In re Oetiker, 977 F.2d 1443, 1445, 24 USPQ2d 1443, 1444 (Fed. Cir. 1992).

Factors to be considered in determining that claims are not obvious include unexpected results, new features, solution of a different problem and novel properties (see In re Wright,848 F.2d 1216, 6 U.S.P.Q.2d 1959(Fed.Cir.1988)).

Hindsight cannot be used in determining the issue of obviousness and the reviewer must view the prior art without reading into that art the teachings of the application or patent (see Kalman v. Kimberly Clark Corp.,713 F.2d 760,218 U.S.P.Q. 781(Fed.Cir.1983)).

"[T]he best defense against the subtle but powerful attraction of a hindsight-based obviousness analysis is rigorous application of the requirement for a showing of the teaching or motivation to combine prior art references . . . . Combining prior art references without evidence of such a suggestion, teaching, or motivation simply takes the inventor's disclosure as a blueprint for piecing together the prior art to defeat patentability--the essence of hindsight." In re Dembiczak, 175 F.3d 994, 999, 50 USPQ2d 1614, 1617 (Fed. Cir. 1999).

To imbue one of ordinary skill in the art with knowledge of the invention in suit, when no prior art reference or references of record convey or suggest that knowledge, is to fall victim to the insidious effect of a hindsight syndrome wherein that which only the inventor taught is used against its teacher (see W.L. Gore & Associates, Inc. v. Garlock, Inc., 220 USPQ 303, 312-13 (Fed. Cir. 1983), cert. denied, 469 U.S. 851 (1984))

According to 37 C.F.R. § 1.56.b.2.ii (emphasis added):

> A prima facie case of unpatentability is established when the information compels a conclusion that a claim is unpatentable under *the preponderance of evidence*, burden-of-proof standard, giving each term in the claim its broadest reasonable construction consistent with the specification, and before any consideration is given to evidence which may be submitted in an attempt to establish a contrary conclusion of patentability.

According to Black's Law Dictionary (5th ed.), a "preponderance of evidence" is "Evidence which is of greater weight or more convincing than the evidence which is offered in opposition to it; that is, evidence which as a whole shows that the fact sought to be proved is more probable than not."

B. The Office Action

On pages 2-8 of the final Office Action, dated September 9, 2005, the Examiner rejected claims 1-29, 44-48, and 50-57 under 35 U.S.C. § 103(a) as being unpatentable over U.S. Patent 6,240,409 to Aiken ("Aiken").

C. The Prior Art

U.S. Patent 6,240,409 to Aiken is directed to a method and apparatus for detecting and summarizing document similarities within large document sets. Specifically, Aiken seeks to determine similarities in a manner that guarantees that if a sub-string of a predetermined length appears in more than one document in the document set, the sub-string will be detected. *See* Aiken, column 2, lines 35-44. A query input file is thus segmented into multiple query file sub-strings. A query file sub-string is then selected and compared to an index file containing multiple ordered file sub-strings taken from previously analyzed files. *See* Aiken, column 2, lines 47- 67.

The Examiner cited Figures 1a and 1b, and column 3, lines 44-47 of Aiken in the Office Action as disclosing a method of detecting similar documents. *See* Office Action, page 3, lines 1-3. Aiken disclosed therein a method of hashing, comparing and storing a query document against documents already stored in an index file. Additionally, the Examiner cited column 6, lines 7 through 28 in Aiken as disclosing a tuple generating step. *See* Office Action, page 3, lines 3-5. Aiken disclosed therein storage of hash value, position pairs into a temporary file. *See* Aiken, column 6, lines 7-10. The position data stored in the temporary file is described as encoding the name of the document and an offset within the file as to where the hash value begins. *See* Aiken, column 6, lines 16-19.

The Examiner cited the detection of similar documents if the tuples of such documents are clustered together as being disclosed by Aiken in Figures 4a, 4b, and 4c; column 7, lines 25-35 and column 10, line 4 - column 12, line2. *See* Office Action, page 3, line 7-9. Aiken disclosed therein integrating a current document into an existing cluster if the Aiken invention determines that the current document has a sufficient number of matches with other previously loaded documents. *See* Aiken, column 7, lines 25-30. The clustering described in Aiken occurs via a find/union operation. *See* Aiken, column 7, lines 31-32.

The Examiner cited the claimed limitation "tokens being eliminated based on parts of speech" as being disclosed by Aiken in column 4, lines 57-57 and column 8 line 67 - column 9, line 3. *See* Office Action, page 3, lines 9-11. Aiken disclosed therein removal of unimportant and/or frequently occurring words from a data string. *See* Aiken column 4, lines 54-58, and column 9, lines 2-3.

The Examiner took official notice that different operating systems use different token ordering. *See* Office Action, page 3, lines 13-14. The Examiner concludes it would have been obvious "to include sorting filtered documents to reorder the tokens according to a predetermined ranking in order to accommodate different operating systems while implementing the method of Aiken." Office Action, page 3, lines 16-18.

D. The Present Claimed Invention Patentably Distinguishes Over The Prior Art

Independent claim 50 recites "filtering the document to eliminate tokens based on parts of speech." Independent claim 51 recites "a filter to filter the document to eliminate tokens based on parts of speech." Independent claim 1 recites "tokens being eliminated based on at least one of (a) parts of speech and (b) collection statistics."

MPEP §2101.01 directs an examiner: "During examination, the claims must be interpreted as broadly as their terms reasonably allow." The Examiner failed to give the words in the claims the broadest meaning such terms would allow. As described by the Examiner, "The claimed 'tokens being eliminated based on parts of speech' is met by the fact that the method of Aiken eliminates stop word[s]." Thus, the Examiner failed to consider grammatical parts of speech by equating "parts of speech" with stop words.

The present invention does not seek to limit the plain meaning of the phrase "parts of speech." In fact, the specification provides examples of various parts of speech that could be filtered out of a token stream. Examples of the parts of speech used to filter the token stream disclosed in one embodiment include nouns, verbs, adjectives, adverbs, prepositions, and types of nouns. *See* specification lines 22-30. Additionally, the present invention explicitly distinguishes token removal based on parts of speech and token removal based on stop words. *Compare* specification page 12, lines 1-6 *to* specification, page 12, lines 22-30.

Based on the above discussion, it is submitted that the Examiner failed to make a *prima facie* showing of obviousness based on the Aiken reference. Aiken fell short of teaching or suggesting the removal of tokens based on parts of speech. While Aiken does teach the removal of stop words from a data string, such a disclosure fell short of the plain meaning, implied and relied on by the present invention, of the phrase "parts of speech."

It is therefore submitted that the Examiner improperly relied on Aiken, which disclosed the removal of stop words from a query string, to disclose the claimed feature "the token being eliminated based on . . . parts of speech."

Because of the "parts of speech" deficiency in the Examiner's Rejection, independent claims 50 and 51 patentably distinguish over the prior art. Withdrawal of the Examiner's rejection is respectfully requested.

With regard to independent claim 1, this claim indicates that tokens are eliminated based on at least one of (a) parts of speech and (b) collection statistics relating to a number of occurrences or words or phrases in the document. As described above, the prior art does not suggest tokens being eliminated based on parts of speech. As discussed below, the prior art does not suggest tokens being eliminated based on collection statistics relating to a number of occurrences of words or phrases in the document.

Aiken failed to disclose "tokens being eliminated based on . . . collection statistics relating to a number of occurrences of words or phrases in the document." The Examiner did not explicitly address this feature of the present invention in the Office Action, dated September 9, 2005. The Examiner cited Aiken, column 9, lines 1- 3, which disclosed removing words used frequently. Applicants thus assumes the Examiner relied on that citation to disclose tokens being eliminated based on collection statistics.

The citation relied on by the Examiner does not fully teach or suggest removal of token based on collection statistics, as disclosed in the present invention. One embodiment of the collection statistics disclosed by the specification contains frequency data of tokens in the document. *See* specification, page 12, lines 18-21. Thus, according to one embodiment, a token may be eliminated based on the frequency of that token in the document as opposed to a pre-determined list of frequently occurring words. Aiken specifically disclosed "removing words 'this' and 'is' under the assumption that they are words that would be used frequently." Aiken does not teach or suggest whether the frequency of a word in a document would be used to remove such a word. Aiken merely suggests that a pre-determined list of frequently used words would be removed from a data string. In the system of Aiken, it is somewhat irrelevant whether the word is frequently used in the document under consideration.

It is therefore submitted that the Examiner improperly relied on Aiken, which discloses the removal of “frequently used” words based on a pre-determined list of such words, to disclose the claimed feature “the token being eliminated based on . . . collection statistics relating to the number of occurrences of words or phrases in the document.”

In summary, it is submitted that the prior art does not teach or suggest a method for similar document detection comprising:

> . . . filtering the document to eliminate tokens and obtain a filtered document containing remaining tokens, the tokens being eliminated based on at least one of (a) parts of speech and (b) collection statistics relating to a number of occurrences of words or phrases in the document;

as set forth in claim 1. Therefore claim 1 is patentably distinguishable over the prior art. Claims 2-28, 44-48, and 52-57 depend directly or indirectly from independent claim 1 and include all features of the claim(s) from which they depend. Therefore, it is submitted that claims 2-28, 44-48, and 52-57 patentably distinguish over the prior art.

Applicants solicit of claims 1-28, 44-48 and 52-57 under 35 U.S.C. § 103(a) because Aiken fails to teach or suggest the above-identified features. Withdrawal of the Examiner's rejection is respectfully requested.

VIII. CONCLUSION

In summary, it is submitted that claims 1-28, 44-48, and 50-57 patentably distinguish over the prior art. Reversal of the outstanding rejections is respectfully requested.

* * *

The Commissioner is authorized to charge any additional Appeal Brief fee or Petition for Extension of Time fee for underpayment or credit any overpayment to Deposit Account 19-3935.

Respectfully submitted,
STAAS & HALSEY LLP

Date: Jul 27 2006

By: ______________________
Mark J Henry
Registration No. 36,162

1201 New York Ave, N.W., Suite 700
Washington, D.C. 20005
Telephone: (202) 434-1500
Facsimile: (202) 434-1501

IX. CLAIMS APPENDIX (37 C.F.R. § 41.67(c)(I)(viii)

1. (previously presented) A method for detecting similar documents comprising the steps of:

obtaining a document;

filtering the document to eliminate tokens and obtain a filtered document containing remaining tokens, the tokens being eliminated based on at least one of (a) parts of speech and (b) collection statistics relating to a number of occurrences of words or phrases in the document;

sorting the filtered document to reorder the tokens according to a predetermined ranking;

generating a single tuple for the filtered document;

comparing the tuple for the filtered document with a document storage structure comprising a plurality of tuples, each tuple in the plurality of tuples representing one of a plurality of documents; and

determining if the tuple for the filtered document is clustered with another tuple in the document storage structure, thereby detecting if the document is similar to another document represented by the another tuple in the document storage structure.

2. (Original) A method as in claim 1, wherein the step of filtering comprises parsing the document, and wherein the filtered document comprises a token stream, the token stream comprising a plurality of tokens.

3. (Original) A method as in claim 2, wherein the step of filtering further comprises retaining a token in the token stream as a retained token according to at least one token threshold.

4. (previously presented) A method as in claim 3, wherein the step of filtering further comprises reordering the retained tokens in the token stream to obtain a reordered token stream.

5. (previously presented) A method as in claim 44, wherein

the step of filtering further comprises retaining a token in the token stream as a retained token according to at least one token threshold; and

the step of determining the hash value for the filtered document comprises determining the hash value by processing individually each retained token in the token stream.

6. (Original) A method as in claim 2, wherein the step of filtering further comprises: determining a score for each token in the token stream;

comparing the score for each token to a first token threshold; and

modifying the token stream by removing each token having a score not satisfying the first token threshold and retaining each token as a retained token having a score satisfying the first token threshold.

7. (Original) A method as in claim 6, wherein the step of filtering further comprises:

comparing the score for each retained token to a second token threshold; and

modifying the token stream by removing each retained token having a score not satisfying the second token threshold and retaining each retained token having a score satisfying the second token threshold.

8. (Original) A method as in claim 2, wherein the step of filtering further comprises removing from the token stream at least one token corresponding to a stop word.

9. (Original) A method as in claim 2, wherein the step of filtering further comprises removing a token from the token stream if the token is a duplicate of another token in the token stream.

10. (previously presented) A method as in claim 2, wherein the step of filtering further comprises removing a token from the token stream based on collection statistics and at least one token threshold.

11. (Original) A method as in claim 2, wherein the step of filtering comprises removing at least one token from the token stream.

12. (Original) A method as in claim 1, wherein the step of filtering comprises removing formatting from the document.

13. (Original) A method as in claim 1, wherein the step of filtering uses collection

statistics for filtering the document.

14. (Original) A method as in claim 13, wherein the collection statistics pertain to the plurality of documents.

15. (previously presented) A method as in claim 44, wherein the step of determining the hash value for the filtered document comprises using a hash algorithm to determine the hash value, the hash algorithm having an approximately even distribution of hash values.

16. (previously presented) A method as in claim 44, wherein the step of determining the hash value for the filtered document comprises using a standard hash algorithm to determine the hash value.

17. (previously presented) A method as in claim 44, wherein the step of determining the hash value for the filtered document comprises using a secure hash algorithm to determine the hash value.

18. (previously presented) A method as in claim 44, wherein the step of determining the hash value for the filtered document comprises using hash algorithm SHA-1 to determine the hash value.

19. (previously presented) A method as in claim 44, wherein the document storage structure comprises a hash table.

20. (Original) A method as in claim 1, wherein the document storage structure comprises a tree.

21. (Original) A method as in claim 20, wherein the tree comprises a binary tree.

22. (Original) A method as in claim 21, wherein the binary tree comprises a binary balanced tree.

23. (Original) A method as in claim 1, wherein the document storage structure comprises

a hash table and at least one tree.

24. (Original) A method as in claim 1, wherein the step of comparing comprises inserting the tuple into the document storage structure.

25. (previously presented) A method as in claim 44,

wherein the document storage structure comprises a hash table, the hash table comprising a plurality of bins, each bin of the hash table comprising at least one tuple of the plurality of tuples, and

wherein the step of determining if the tuple is clustered with another tuple comprises determining if the tuple is co-located with another tuple at a bin of the hash table.

26. (Original) A method as in claim 1, wherein the document storage structure comprises a tree, the tree comprising a plurality of branches, each bucket of the tree comprising at least one tuple of the plurality of tuples, and

wherein the step of determining if the tuple is clustered with another tuple comprises determining if the tuple is co-located with another tuple in a bucket of the tree.

27. (Original) A computer for performing the method of claim 1.

28. (Original) A computer-readable medium having software for performing the method of claim 1.

29. (cancelled)

30. (previously presented) A method for detecting similar documents comprising the steps of:

obtaining a document;

parsing the document to remove formatting and to obtain a token stream, the token stream comprising a plurality of tokens;

retaining only retained tokens in the token stream by using at least one token threshold;

reordering the retained tokens to obtain an arranged token stream;

processing in turn each retained token in the arranged token stream using a hash algorithm to obtain a single hash value for the document;

generating a document identifier for the document;

forming a single tuple for the document, the tuple comprising the document identifier for the document and the hash value for the document;

inserting the tuple for the document into a document storage tree, the document storage tree comprising a plurality of tuples, each tuple located at a bucket of the document storage tree, each tuple in the plurality of tuples representing one of a plurality of documents, each tuple in the plurality of tuples comprising a document identifier and a hash value; and

determining if the tuple for the document is co-located with another tuple at a same bucket in the document storage tree, thereby detecting if the document is similar to another document represented by the another tuple in the document storage tree.

31. (Original) A computer for performing the method of claim 30.

32. (Original) A computer-readable medium having software for performing the method of claim 30.

33. (previously presented) An apparatus for detecting similar documents comprising: means for obtaining a document;

means for parsing the document to remove formatting and to obtain a token stream, the token stream comprising a plurality of tokens;

means for retaining only retained tokens in the token stream by using at least one token threshold;

means for reordering the retained tokens to obtain an arranged token stream;

means for processing in turn each retained token in the arranged token stream using a hash algorithm to obtain a single hash value for the document;

means for generating a document identifier for the document;

means for forming a single tuple for the document, the tuple comprising the document identifier for the document and the hash value for the document;

means for inserting the tuple for the document into a document storage tree, the document storage tree comprising a plurality of tuples, each tuple located at a bucket of the document storage tree, each tuple in the plurality of tuples representing one of a plurality of documents, each tuple in the plurality of tuples comprising a document identifier and a hash value; and

means for determining if the tuple for the document is co-located with another tuple at a same bucket in the document storage tree, thereby detecting if the document is similar to another document represented by the another tuple in the document storage tree.

34 - 43 (cancelled)

44. (previously presented) A method as claimed in claim 1, wherein

the method further comprises determining a document identifier for the filtered document and a single hash value for the filtered document,

the tuple comprises the document identifier for the filtered document and the hash value for the filtered document, and

each tuple in the plurality of tuples comprising a document identifier and a hash value.

45. (previously presented) A method as claimed in claim 1, wherein filtration is based on parts of speech.

46. (previously presented) A method as claimed in claim 1, wherein filtration removes frequently occurring terms.

47. (previously presented) A method as claimed in claim 1, wherein filtration removes infrequently occurring terms.

48. (previously presented) A method as claimed in claim 1, wherein filtration eliminates words having an occurrence frequency that falls within a pre-determined frequency range.

49. (previously presented) A method as claimed in claim 30, wherein reordering is based on Unicode ordering.

50. (previously presented) A method for detecting similar documents comprising the steps of:

obtaining a document;

filtering the document to eliminate tokens based on parts of speech and obtain a filtered document;

generating a single tuple for the filtered document;

comparing the tuple for the filtered document with a document storage structure comprising a plurality of tuples, each tuple in the plurality of tuples representing one of a plurality of documents; and

determining if the tuple for the filtered document is clustered with another tuple in the document storage structure, thereby detecting if the document is similar to another document represented by the another tuple in the document storage structure.

51. (previously presented) An apparatus for detecting similar documents comprising:

means for obtaining a document;

a filter to filter the document to eliminate tokens based on parts of speech and obtain a filtered document;

a tuple unit to generate a single tuple for the filtered document;

a comparator to compare the tuple for the filtered document with a document storage structure comprising a plurality of tuples, each tuple in the plurality of tuples representing one of a plurality of documents; and

a decision unit to determine if the tuple for the filtered document is clustered with another tuple in the document storage structure, based on the comparison, thereby detecting if the document is similar to another document represented by the another tuple in the document storage structure.

52. (previously presented) A method as in claim 3, wherein the token threshold represents tokens that are frequently used in the document collection.

53. (previously presented) A method as in claim 52, wherein frequently used is determined by inverted document frequency scores.

54. (previously presented) A method as in claim 3, wherein the token threshold represents tokens that are infrequently used in the document collection.

55. (previously presented) A method as in claim 54, wherein frequently used is determined by inverted document frequency scores.

56. (previously presented) A method as in claim 3, wherein upper and lower bound token thresholds represent tokens that are within a range of frequency of use in the document collection.

57. (previously presented) A method as in claim 56, wherein frequency of use is determined by inverted document frequency scores.

58. (previously presented) A method as claimed in claim 30, wherein reordering is based on Unicode ordering.

59. (previously presented) A method as claimed in claim 30, wherein reordering is based on EBCDIC ordering.

60. (previously presented) A method as claimed in claim 30, wherein reordering is based on ASCII ordering.

61. (previously presented) A method as claimed in claim 30, wherein reordering is based on collection statistic measurements.

62. (previously presented) A method as claimed in claim 61, wherein collection statistic measurements are determined based on an inverse document frequency.

X. EVIDENCE APPENDIX (37 C.F.R. § 41.337(c)(l)(ix))

Not applicable.

XI. RELATED PROCEEDING APPENDIX (37 C.F.R. § 41.337(c)(l)(x))

Not applicable.